(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
11 March 2004 (11.03.2004)

PCT

(10) International Publication Number
WO 2004/021669 A1

(51) International Patent Classification[7]: H04L 29/06, 12/56

(21) International Application Number:
PCT/SG2002/000203

(22) International Filing Date:
2 September 2002 (02.09.2002)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant *(for all designated States except US)*: INFINEON TECHNOLOGIES AG [DE/DE]; St.-Martin-Strasse 53, 81669 Munich (DE).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: MISHRA, Shridhar, Mubaraq [IN/US]; 1325A Spruce Street, Berkeley, CA-94709 (US). ARDHANARI, Guruprasad [IN/SG]; Blk 109 Hougang Ave 1 #09-1012, Singapore 530109 (SG).

(74) Agent: WATKIN, Timothy, Lawrence, Harvey; Lloyd Wise, Tanjong Pagar, P.O. Box 636, Singapore 910816 (SG).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

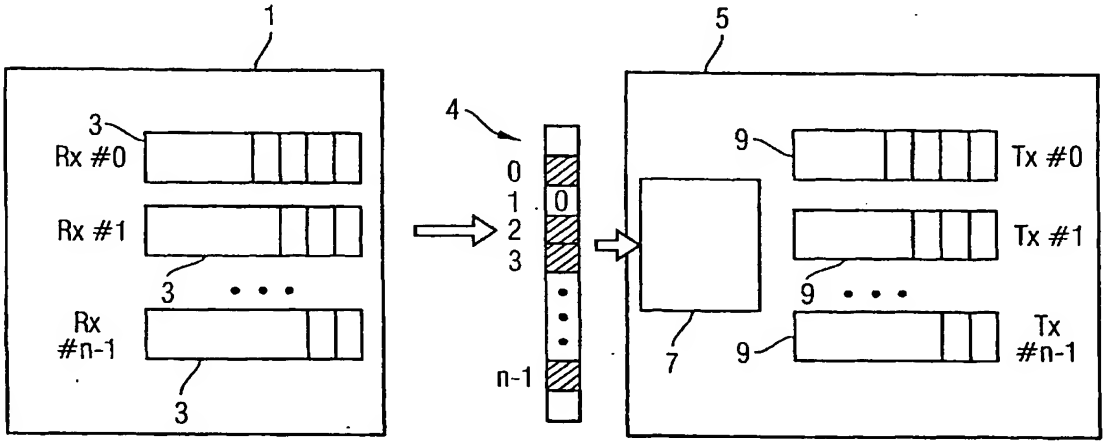
(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declaration under Rule 4.17:
— *of inventorship (Rule 4.17(iv)) for US only*

Published:
— *with international search report*
— *with amended claims*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: A DATA SWITCH AND A METHOD FOR BROADCAST PACKET QUEUE ESTIMATION



(57) Abstract: A data switch includes ingress ports associated with ingress queues (3) and egress ports associated with egress queues (9). The length of the ingress queues (3) is measured, and the level of broadcast packets arriving at the ingress ports is thereby estimated. Based on this estimate it is determined whether or not the level of broadcast packets is excessive, and in this case broadcast storm control is carried out.

WO 2004/021669 A1

A DATA SWITCH AND A METHOD FOR BROADCAST PACKET QUEUE ESTIMATION

Field of the invention

The present invention relates to a data switch and to a method of operating it.

Background of Invention

One of the types of data packets which Ethernet switches are required to transmit are broadcast packets, i.e. packets which are to be transmitted from one of the ingress ports to all of the egress ports, except the egress port corresponding to the ingress port ("source port") from which the broadcast packet arrived. Shared memory output queue Ethernet switches cannot sustain excessive levels of broadcast packets, because the memory requirements increase linearly with the percentage of broadcasts in a traffic stream. This means that there is a need to limit the number of broadcasts in the system.

In the case that it is identified that the number of broadcast packets is excessive, it is known to delete selected ones of the broadcast packets, e.g. selectively based on a parameter in the header of the packet defining the importance of the packet. This is referred to as "broadcast storm control" (BSC).

Conventional methods to identify excessive amounts of broadcast packet traffic operate by counting the number of broadcasts per unit time. Once this value rises above a predefined level, BSC is turned on. When the figure drops below the predetermined level (e.g. by a certain amount, so that there is a hysteresis), BSC is turned off. This method suffers from the problem that it requires a counter for explicitly counting the broadcast packets. Additionally, since the count must be worked out per unit time, a timer is required, e.g. to decrement the counter every timer interval.

Summary of the Invention

The present invention proposes a new and useful manner of determining excess levels of broadcast packets, in particular so that BSC can be carried out.

In general terms the invention proposes that the length of the respective queues at the ingress ports is measured, and the level of broadcast packets is estimated, or in some circumstances exactly determined, based on these lengths. The method is motivated by the observation that a broadcast packet takes longer than a normal packet to pass through the switch, and therefore causes the length of the queue to grow. In wirespeed unicast systems with one-to-one traffic flow, broadcast packets are in fact the only types of packets which can cause the ingress queues to lengthen.

From the level of the broadcast packets, a determination is made of whether or not the level is excessive, and in this case BSC can be carried out, for example according to the conventional methods described above. For example, BSC can be carried out whenever the system determines that the length of any of the queues rises above a predetermined level, since the length of that queue provides a measure of the frequency of arrival of broadcast packets (at the corresponding ingress port).

Brief Description of The Figures

Preferred features of the invention will now be described, for the sake of illustration only, with reference to Fig. 1, which shows schematically a switch according to the invention.

Detailed Description of the embodiments

Referring to Fig. 1, a Ethernet switch which is an embodiment of the invention is shown. According to conventional structures, the Ethernet switch has a number of ingress ports n and a corresponding number n of egress ports. Data packets arrive at the ingress ports for transmission across a switching fabric to the egress ports.

The Ethernet switch has a packet resolution module 1 including a respective ingress queue 3 for each ingress port. The ingress queues are marked from Rx#0 up to Rx#n-1. The packet resolution module 1 determines a destination list for each packet arriving at a certain ingress port (i.e. a list of the egress ports to which it should be transmitted), and stores this information in the corresponding queue. The destination list for a typical packet is labelled 4 in Fig. 1, and includes for each of the n destinations either an indication that the packet is to be sent there (marked in destination list 4 as a black square), or that it is not (marked as a 0). The destination list 4 shown in Fig. 1 is for a broadcast packet having ingress 1 as the source port, so that it is 0 for destination 1, and a black square for all other destinations.

The Ethernet switch further includes a queue management module 5 having a scheduler 7 and a respective egress queue 9 for each of the n egress ports. The egress queues are marked from Tx#0 up to Tx#n-1. The scheduler 7 in the queue management module 5 processes packets from each ingress port in a round-robin manner. For each packet the packet details are transmitted into all the egress queues specified in the destination list for that packet. The time taken for this insertion depends upon the amount of parallelism available in the queue management module 5, and is referred to as the scheduler bandwidth, which may be 5 insertions per unit time.

Each of the broadcast packets have to be inserted into each of the egress queues (except the source port), so if a broadcast packet arrives in the ingress queue structure every unit time, the scheduler must have a bandwidth

of n-1 to match the ingress bandwidth (even in the absence of other packets). If the scheduler bandwidth is less than this, the ingress queue sizes will increase.

Specifically, suppose that the packet rate at each ingress port is *M* packets per unit time ($0 \le M \le 1$), so that the total number of packets arriving at the switch per unit time is NM. Suppose that the broadcast traffic as a fraction of all traffic is *b* ($0 \le b \le 1$), and that the actual scheduler bandwidth is S per unit time. In this case, the required scheduler rate is $NM(1-b)+bNM(N-1)$ which is equal to $NM(1+(N-2)b)$ per unit time. The difference between the egress and ingress rates is thus $NM(1-b)+bNM(N-1)-S$, and the rate of increase of the ingress queues is therefore $\{\ NM(1-b)+bNM(N-1)-S\ \}/N$.

In the embodiment, the packet resolution module 1 is arranged to determine the length of each of the queues, and according to the lengths determine if BSC should be applied. Preferably, the packet resolution module determines that this is the case when it finds that the length of any one of the queues rises above a predetermined level. Alternatively (or additionally), the packet resolution module may determine that this is the case when it finds that the total length of the n queues (i.e. the sum of the lengths of the n queues) rises above this predetermined maximum.

Once BSC has been applied, the packet resolution module 1 continuously monitors whether it must be turned off again. For example, if the BSC was triggered by the length of any one of the queues rising above a predetermined level, the BSC may be removed again in the case that it is found that the length of that queue has now fallen below a second predetermined level. Similarly, in the case that BSC was triggered by the total length of the queues rising above the predetermined level, the BSC may be removed in the case

that it is found that the total length of the queues falls below a second predetermined level. In either case, the second predetermined level must be no higher than the first predetermined level, and is preferably lower since this provides a hysteresis.

Although only a single embodiment of the method has been described above, the invention is not limited in this respect and many variations are possible, just as there are many known designs of Ethernet switch. In particular, different Ethernet switches manage their ingress ports in different manners, but the general principle of measuring the lengths of ingress queues and obtaining from them a measure of the proportion of broadcast packets remains valid.

Claims

1. A data switch having a plurality of ingress ports and egress ports connected by a switching fabric, the switch having a plurality of ingress queues for queuing data derived from data packets arriving at the ingress ports, the switch further comprising broadcast packet estimation means for deriving a measure of the length of at least one of the queues and using it to obtain a measure of the frequency of arrival of broadcast packets.

2. A data switch according to claim 1 in which the broadcast packet estimation means determines the measure of the frequency of arrival of broadcast packets as the length of the longest of the queues.

3. A data switch according to claim 1 or claim 2 further including a broadcast packet control means for deleting at least some of the broadcast packets when the broadcast packet estimation means indicates that the measure of the frequency of arrival of broadcast packets is above a first predetermined level.

4. A data switch according to claim 3 in which the broadcast packet deletion means is arranged to cease deleting packets when the broadcast packet estimation means indicates that the measure of the frequency of arrival of broadcast packets is below a second predetermined level.

5. A method of operating a data switch having a plurality of ingress ports and egress ports connected by a switching fabric, the switch having a plurality of ingress queues for queuing data derived from data packets arriving at the ingress ports, the method comprising:

deriving a measure of the length of at least one of the queues; and

using the measure of the length of at least one of the queues to obtain a measure of the frequency of arrival of broadcast packets.

6. A method according to claim 5 in which the measure of the frequency of arrival of broadcast packets is the length of the longest of the queues.

7. A method of according to claim 5 or claim 6 further including, when the measure of the frequency of arrival of broadcast packets rises above a first predetermined level, commencing deleting at least some of the broadcast packets.

8. A method according to claim 7 further including ceasing to delete packets when the measure of the frequency of arrival of broadcast packets falls below a second predetermined level.

AMENDED CLAIMS
[received by the International Bureau on 20 October 2003 (20.10.03);
original claims 1-8 replaced by new claims 1-8 (2 pages)]

Claims

1. A data switch having a plurality of ingress ports and egress ports connected by a switching fabric, the switch having a plurality of ingress queues for queuing data derived from data packets arriving at the ingress ports, the switch being characterised by further comprising: broadcast packet estimation means for deriving a measure of the length of at least one of the queues and using it to obtain a measure of the frequency of arrival of broadcast packets; and a broadcast packet control means arranged to be triggered according to the measure of the frequency of arrival of broadcast packets into a broadcast storm control mode in which the broadcast packet control means performs broadcast storm control.

2. A data switch according to claim 1 in which the broadcast packet estimation means determines the measure of the frequency of arrival of broadcast packets as the length of the longest of the queues.

3. A data switch according to claim 1 or claim 2 in which the broadcast packet control means is arranged to perform the broadcast storm control by deleting at least some of the broadcast packets when the broadcast packet estimation means indicates that the measure of the frequency of arrival of broadcast packets is above a first predetermined level.

4. A data switch according to claim 3 in which the broadcast packet deletion means is arranged to cease deleting packets when the broadcast packet estimation means indicates that the measure of the frequency of arrival of broadcast packets is below a second predetermined level.

5. A method of operating a data switch having a plurality of ingress ports and egress ports connected by a switching fabric, the switch having a plurality of ingress queues for queuing data derived from data packets arriving at the ingress ports, the method comprising:

deriving a measure of the length of at least one of the queues;

and characterised by:

using the measure of the length of at least one of the queues to obtain a measure of the frequency of arrival of broadcast packets; and

according to the measure of the frequency of arrival of broadcast packets triggering a broadcast storm control mode in which broadcast storm control is performed.

6. A method according to claim 5 in which the measure of the frequency of arrival of broadcast packets is the length of the longest of the queues.

7. A method of according to claim 5 or claim 6 in which the broadcast storm control mode is triggered when the measure of the frequency of arrival of broadcast packets rises above a first predetermined level, and the broadcast storm control is by deleting at least some of the broadcast packets.

8. A method according to claim 7 further including ceasing to delete packets when the measure of the frequency of arrival of broadcast packets falls below a second predetermined level.

FIG 1